Regresi Spline untuk Pemodelan Bidang Kesehatan: Studi tentang Knot dan Selang Kepercayaan

Spline Regression Modelling for Health Problem: Study of Knot and Confidence Interval

Netty Herawati
Jurusan Matematika FMIPA Universitas Lampung

ABSTRACT

This article aimed to study about knot and confidence interval for health science using spline nonparameteric regression. The study used simulation and real data. The result showed that numbers of knot should be placed according to the quantil variable in order to get a good estimation of the data function. In additon, confidence interval using bayesian and bootstrap method gave no different result for a small sample size whereas for a big sample size bootstrap gave narrower interval.

Keywords : Knot, confidence interval, spline regression

PENDAHULUAN

Analisis regresi merupakan salah satu alat statistik yang banyak digunakan untuk mengetahui hubungan antara dua variabel atau lebih. Misalkan diberikan data $\{(x_i , y_i)\ i = 1,2, ...,n\}$, $t_i \in \mathrm{R}$, $y_i \in \mathrm{R}$ dan hubungan antara t_i dan y_i diasumsikan mengikuti model regresi

$$y_i = f(x_i) + \varepsilon_i, t_i \in [a,b],\ i = 1,2,\ldots,n \qquad (1)$$

dengan $f(x_i)$ adalah fungsi regresi dan ε_i adalah sesatan acak yang diasumsikan berdistribusi normal, independen dengan nilai tengah nol dan ragam σ^2.

Ada dua metode yang dapat digunakan untuk menaksir fungsi $f(x_i)$, yaitu metode regresi parametrik dan metode regresi nonparametrik. Metode regresi parametrik akan sesuai jika bentuk fungsi $f(x_i)$ diketahui. Tetapi jika fungsi $f(x_i)$ tersebut tidak diketahui bentuknya, maka metode regresi nonparametrik lebih sesuai digunakan. Dalam hal ini fungsi $f(x_i)$ hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi tersebut biasanya dimotivasi oleh sifat kemulusan (*smoothness*) yang dimiliki oleh fungsi $f(x_i)$ tersebut. Beberapa pendekatan nonparametrik yang cukup populer dalam mengestimasi fungsi $f(x_i)$ antara lain Spline (Wahba 1990, Takezawa 2006), dan penduga *Kernel* (Hardle 1990).

Pada pendekatan non-parametrik, pengepasan (*fitting*) kurva regresi dilakukan dengan memperhatikan peubah respon Y secara terbatas di sekitar x pada selang tertentu, tidak pada keseluruhan pengamatan x. Pada penduga kernel fungsi $f(x_i)$ dimuluskan dengan menggunakan pembobotan terhadap variabel respon Y di sekitar x. Untuk memuluskan $f(x_i)$ perlu dilakukan pemilihan bandwidth yang optimal. Pada Spline pendekatan dilakukan pada segmentasi x untuk membangun fungsi $f(x_i)$ dengan membagi pengamatan x berdasarkan titik-titik x yang disebut knot. Pendekatan ini merupakan *piecewise polynomial*, yaitu polinomial yang memiliki sifat tersegmen pada selang x yang terbentuk oleh titik-titik knot (Wang & Yang 2009). Fungsi $f(x_i)$ kemudan diduga secara lokal pada selang-selang tersebut, dan kemudian diinterpolasi sepanjang keseluruhan pengamatan x dengan pendekatan kuadrat terkecil yang terpenalti (*Penalized Least Square*). Penalti yang digunakan adalah penalti pada kekasaran/kemulusan fungsi dugaan yang diinginkan.

Tulisan ini bertujuan untuk menentukan knot dan penempatannya serta selang kepercayaan bayes dan bootstrap dengan regresi spline pada pemodelan data kesehatan.

Spline

Spline adalah potongan polinomial order r. Titik bersama dari potongan-potongan tersebut disebut dengan knots. Spline order r dengan knots pada $\zeta_1,\ldots,\zeta_k$ diberikan dalam fungsi S dengan bentuk (Eubank 1988, Schumaker 2007).

$$S(x) = \sum_{i=0}^{r-1} \theta_i x^i + \sum_{j=1}^{k} \delta_j \, (x-\zeta_j)_+^{r-1} \qquad (2)$$

$$\text{dan } (x-\zeta_j)_+^{r-1} = \begin{cases} (x-\zeta_j) & ,(x-\zeta_j) \geq 0 \\ 0 & ,(x-\zeta_j) < 0 \end{cases}$$

Spline mempunyai sifat :
S merupakan potongan polinomial derajat *r – 1* pada setiap subinterval $[\zeta_j, \zeta_{j+1}]$. *S* mempunyai turunan ke (*r*-2) yang kontinyu. *S* mempunyai turunan ke (*r*-1) yang merupakan fungsi tangga dengan titik-titik lompatan pada $(\zeta_1, ..., \zeta_k)$. Apabila didefinisikan suatu spline alami berorde *r* = 2*m* dengan titik-titik knots pada $x_1, \ldots, x_n$ yaitu spline yang memenuhi sifat 1, 2, dan 3 juga memenuhi *S* adalah polynomial derajat *m*-1 diluar interval [x_1, x_n]. *S* memenuhi syarat batas alami (*natural boundary condition*), yaitu $s^{(j)}(a) = s^{(j)}(b) = 0$, *j* = *m*,…, *2m*-1 (Green & Silverman 1994).

Jika dalam persamaan (2) diambil nilai *r* = 4, maka di peroleh spline kubik yang memenuhi syarat berikut: pada setiap interval (a,x_1) , (x_1 , x_2), …, (x_n, b), *f* adalah polinomial kubik. Turunan pertama dan kedua dari *f* kontinyu pada setiap $x_i \in [a,b]$ dengan x_i titik knots (Green & Silverman 1994).

Spline dalam regresi non-parametrik

Regresi nonparametrik spline, dari fungsi f seperti pada (1) sebagai berikut:

$y_i = f(x_i) + \varepsilon_i, t_i \in [a,b]$, *i* =1,2,…,*n*.

Pemulusan spline akan mengestimasi fungsi f sebagai solusi dari masalah optimasi yaitu dengan mencari $\hat{f} \in L_2[a,b]$ yang meminimumkan jumlah kuadrat galat terpenalti (*penalized residual sum of square*) sebagaimana persamaan (3) berikut:

$$S(f) = n^{-1} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(x)\}^2 dx \qquad (3)$$

Untuk nilai λ>0, dimana $L_2[0,1]$ menyatakan himpunan fungsi-fungsi kuadrat terintegral pada interval [a,b]. Suku pertama persamaan (3) adalah kuadrat tengah sisaan (*Mean Square Error, MSE)*. Sedangkan suku kedua, yang diboboti dengan λ (parameter pemulus), merupakan penalti kekasaran (*roughness penalty*) yang memberikan ukuran kemulusan atau kekasaran kurva dalam memetakan data, melalui parameter penghalus λ≥ 0. Dengan kata lain, suku kedua akan mempenalti kurva dari fungsi *f*. Nilai λ bervariasi dari 0 sampai +∞, jika λ → +∞, maka penalti mendominasi persamaan (3) dan penduga spline menjadi konstan. Sebaliknya, jika λ → 0, maka penalti tidak lagi ada dalam persamaan (3) dan penduga spline menginterpolasi data. Dengan demikian, parameter penghalus λ memainkan peran sentral dalam mengendalikan perimbangan (*trade off*) antara ketepatan model (*goodness of fit*) dan mulusnya penduga.

Solusi yang diperoleh dari pemulusan spline dengan meminimumkan persamaan (3) dikenal sebagai spline kubik (natural cubic spline atau cubic spline) dengan knot pada x1, ..., xn . Dengan sudut pandang ini, interpolasi spline yang bergantung pada pemilihan parameter pemulusan λ memiliki struktur khusus sebagai suatu pendekatan yang cocok dan pas untuk fungsi f dalam model regresi nonparametrik persamaan (1).

Fungsi Spline berorde ke-m adalah sembarang fungsi yang secara umum dapat disajikan dalam bentuk:

$$f(x_i) = \beta_0 + \sum_{j=1}^{p} \left[\sum_{r=1}^{m-1} \beta_{j,r} X_j^r + \sum_{k=1}^{s} \beta_{j,(m-1),k} \left(X_j - K_{jk} \right)_+^{m-1} \right] \qquad (4)$$

dengan fungsi terpancung sebagai berikut:

$$(X_j - K_{jk})_+^{m-1} = \begin{cases} \left(X_j - K_{jk} \right)^{m-1} & ; \quad X_j \geq K_{jk} \\ 0 & ; \quad X_j < K_{jk} \end{cases}$$

di mana:
β = Parameter model.
β0 = Intersep
βjr = Slope pada peubah X_j dengan orde ke-r
β j(m-1)k = Slope pada peubah X_j truncated knot ke-k pada Spline ber-orde m
X_j = Peubah penjelas ke-j
K_{jk} = Knot ke-k pada peubah X_j
J = 1, 2, …, p dan k = 1, 2, …, s_j
s_j = Banyaknya knot dalam peubah penjelas ke-j

Untuk fungsi spline dengan satu peubah penjelas, yakni j = 1, bentuk umumnya:

$$f(x_i) = \beta_0 + \sum_{r=1}^{m-1} \beta_r X^r + \sum_{k=1}^{s_j} \beta_{(m-1),k} \left(X - K_k \right)_+^{m-1}$$

Dari bentuk matematis fungsi Spline tersebut, dapat dikatakan bahwa spline merupakan model polinomial yang tersegmen (piecewise polynomial). Hanya saja, spline justru bersifat kontinu pada knot-knotnya. Knot

diartikan sebagai suatu titik fokus dalam fungsi spline, sehingga kurva yang dibentuk tersegmen pada titik tersebut. Spline orde ke-m, dapat juga diartikan sebagai model polynomial orde ke-m pada tiap interval segmennya, yakni [Kk ,Kk+1]. Hal ini berarti, fungsi Spline merupakan suatu gabungan fungsi polinomial, dimana penggabungan beberapa polinomial tersebut dilakukan dengan suatu cara yang menjamin sifat kontinuitas pada knot-knot. Spline adalah potongan polynomial yang mulus yang masih memungkinkan memiliki sifat tersegmen.

Misalkan $\mathbf{f} = (f(x_1), ..., f(x_n))$ adalah vektor nilai-nilai fungsi f pada titik-titik knot x_1, ..., x_n. Pemulusan spline memberikan $\hat{\mathbf{f}}_\lambda$ sebagai penduga bagi **f** atau nilai dugaan (*fitted value*) bagi $y = (y_1, ..., y_n)^T$ sebagai berikut:

$$\hat{\mathbf{f}}_\lambda = \begin{bmatrix} \hat{f}_\lambda(x_1) \\ \vdots \\ \vdots \\ \hat{f}_\lambda(x_n) \end{bmatrix}_{n\times 1} = \mathbf{A}(\lambda)_{n\times n} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \text{atau} \quad \hat{\mathbf{f}}_\lambda = \mathbf{A}_\lambda \mathbf{y} \tag{5}$$

Dengan $\hat{\mathbf{f}}_\lambda$ adalah spline kubik dengan knot pada x_1, ..., x_n untuk parameter pemulus tertentu. λ>0, dan Aλ adalah matriks pemulus yang simetrik-positif-definit dan tergantung pada λ dan knot x_1, ..., x_n, tetapi bebas dari **y**.

Pemilihan parameter pemulus: metode validasi silang

Untuk menduga bentuk fungsi f , fungsi f diasumsikan *mulus* dan kontinu mutlak pada [a,b] dan $f^{(m)} \in L_2[a,b]\}$. Idealnya akan dipilih suatu nilai λ yang meminimumkan fungsi kerugian $L(\lambda)$, akan tetapi dalam regresi nonparametrik tidak dapat dilakukan secara nyata sebab $L(\lambda)$ masih memuat fungsi f yang tidak diketahui. Sehingga perlu mengestimasi data dan kemudian estimatornya diminimumkan terhadap λ untuk mendapat estimator f yang paling baik (Eubank 1988).

Salah satu pemilihan parameter penghalus λ adalah menggunakan metode validasi silang umum (*Generalized Cross Validation* disingkat GCV) yang merupakan modifikasi dari metode validasi silang (*Cross Validation* disingkat CV) (Green & Silverman 1994). Metode validasi silang umum cukup popular dan disenangi karena tidak memerlukan pengetahuan tentang σ^2, invarian terhadap transformasi (Wahba 1990) dan mempunyai sifat optimal asimptotik. Tujuan dari teori estimasi adalah mencari suatu estimator yang meminimumkan fungsi resiko secara uniform, bila diberikan *n* titik data yang digunakan untuk memilih model, maka akan dibagi menjadi dua bagian yaitu bagian pertama terdiri dari n_x titik data digunakan untuk mencocokkan model, bagian kedua yaitu n-n_x untuk menaksir kemampuan prediksi model.

Tujuan dari teori estimasi adalah mencari suatu estimator yang meminimumkan fungsi resiko secara uniform, keinginan ideal ini sulit untuk diperoleh, sehingga suatu cara untuk mengatasinya adalah membatasi kelas estimator pada estimator linear, yakni estimator yang merupakan fungsi linear observasi. Dari persamaan (1), anggap $C(\Lambda)=\{f_\lambda \mid \lambda \in \Lambda$, Λ = himpunan indeks $\}$ sebagai kelas estimator linear untuk $f(x)$ artinya untuk setiap λ, terdapat matriks $A(\lambda)$ berukuran n×n sehingga :

$$\hat{\mathbf{f}}_\lambda = \mathbf{A}(\lambda)\, y \tag{6}$$

Jika σ^2 diketahui maka λ optimal dapat diperoleh secara langsung dari kriteria prediksi kuadrat tengh galat atau fungsi kerugian yang didefenisikan oleh

$$L(\lambda) = n^{-1}\sum_{i=1}^{n}(f_i - f_{\lambda i})^2 \tag{7}$$

Dalam hal σ^2 tidak diketahui maka dapat digunakan metode validasi silang umum, untuk mendapatkan nilai λ optimal. Metode validasi silang memilih λ yang meminimumkan

$$\mathrm{CV}(\lambda) = n^{-1}\sum_{j=1}^{n}(y_j - f_\lambda^{[j]}(x))^2 \tag{8}$$

$f_\lambda^{[j]}(x)$ adalah f yang meminimumkan $n^{-1}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda\int_0^1 \{f''(x)\}^2 dx$.

Sedangkan metode validasi silang umum memilih λ yang meminimumkan

$$GCV(\lambda) = n^{-1}\sum_{j=1}^{n}\frac{\left(y_j - f_\lambda(t_j)\right)^2}{\left(1 - n^{-1}\sum_{i=1}^{n} a_{jj}(\lambda)\right)^2} \tag{9}$$

Dari persamaan (9) nilai λ yang optimal adalah berkaitan dengan nilai GCV(λ) yang minimum.

Selang kepercayaan penduga spline

Dengan memandang model spline dalam perspektif Bayesian Wahba (1983) mengusulkan selang kepercayaan untuk penduga bagi fungsi regresi spline sebagai:

$$\hat{f}_{\lambda}(x_i) \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2 a_{ii}(\lambda)}$$

Dengan aii(λ) adalah unsur diagonal utama ke-i dari matriks A(λ) hasil validasi silang persamaan (6) dan zα/2 adalah titik dari sebaran normal. Sedangkan $\hat{\sigma}^2$ diperoleh dari kuadrat tengah galat. Selang kepercayaan ini diinterpretasikan sebagai selang kepercayaan bagi seluruh kurva dugaan y dan bukan sebagai selang interval bagi penduga titik. Hal ini dapat dipahami karena fungsi spline yang dihasilkan adalah fungsi sepanjang pengamatan x sebagaimana pada model regresi parametrik, hanya saja ia berupa polinom yang tersegmentasi, dan kemudian dimuluskan sepanjang kurva pula berdasarkan satu nilai penalti pemulus.

Untuk memberikan penduga selang pada titik-titik pengamatan, dapat dilakukan dengan membuat selang kepercayaan bagi dugaan fungsi $y = f(x)$ dengan dua pendekatan. Pertama dengan pendekatan bayesian, yang memberikan selang kepercayaan bayes. Sedangkan pendekatan lain adalah pendekatan bootstrap. Melalui bootstraping pendekatan ini menggunakan ragam penduga empirik dan memberikan selang penduga bagi kurva. Selang yang dihasilkan disebut selang kepercayaan bootstrap. Wang & Wahba (1995) membandingkan beberapa selang kepercayaan Bootstrap dengan selang kepercayaan Bayesian untuk regresi spline. Mereka menyimpulkan bahwa selang kepercayaan Bootstrap sama baiknya dengan interval Bayesian dalam hal rata-rata peluang ketercakupan (coverage probability). Namun selang kepercayaan Bootstrap tampak lebih baik untuk ukuran sample kecil. Tipe interval Bootstrap yang digunakan adalah interval bootstrap tipe "percentile-t interval". Baik selang kepercayaan bayesian maupun booostrap keduanya adalah selang untuk sepanjang kurva, bukan penduga interval bagi setiap titik.

METODE

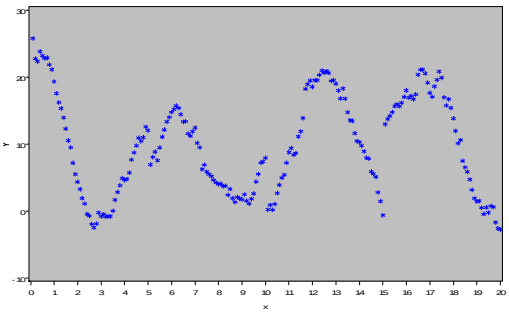
Penelitian dilakukan dengan menggunakan data hasil simulasi dan menggunakan data ril. Simulasi dilakukan untuk mempelajari tentang penempatan dan penentuan jumlah knot yaitu dengan membangkitkan data *y* sebagai peubah respon regresi yang kontinu dan *x* sebagai peubah bebas. Nilai *y* tergantung pada nilai *x* namun tak kontinu pada beberapa titik. Agar memudahkan dan tetap mempertahankan kompleksitas fungsi dipilih fungsi berbasis sinus dengan ketakkontinuan pada titik *x* = {5, 10, 15}. Fungsi yang telah ditetapkan, kemudian diduga dengan regresi non parametrik Spline. Pengepasan regresi dilakukan dengan program SAS, PROC TRANSREG.

Selanjutnya ilustrasi penggunaan regresi spline dilakukan pada data pengukuran kerapatan relatif tulang belakang manusia dari 485 orang di Amerika Utara, tahun 1999 (Hastie *et al*. 2001) dan untuk menduga selang kepercayaan dengan metode bootstrap dan bayesian. Kita akan memodelkan kerapatan relatif ini dengan peubah usia.

HASIL DAN PEMBAHASAN

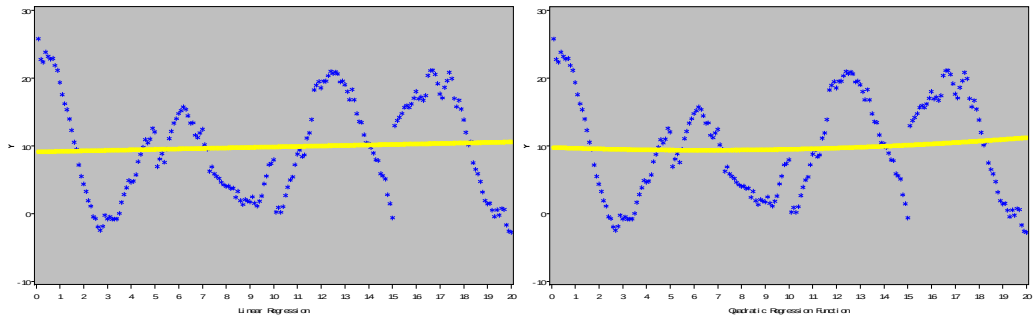
Studi simulasi knot dan penempatannya

Gambar 1 memberikan plot hubungan *x* dan *y* dari data hasil simulasi. Tampak bahwa terdapat ketakkontinuan pada titik-titik *x* =5, 10, dan 15.



Gambar 1. Scater plot *X* dan *Y*. *Y=f(X)* takkontinu pada *x*=5, 10 dan 15**.**

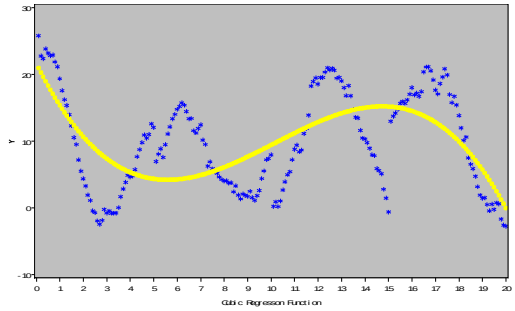
Pada Gambar 2, terlihat bahwa hubungan X dan Y tampak tidak dapat secara sederhana diwakili oleh suatu fungsi regresi parametrik yang didasarkan pada beberapa asumsi. Demikian juga bila kita menggunakan regresi linier atau pun kudratik kita tidak akan memperoleh fungsi baik. Regresi linier hanya



(a) **(b)**

Gambar 2. Plot regresi parametrik (a) linier R^2= 0.27863, (b) polinomial kudratik R^2= 0.46324.

mampu menjelaskan keragaman Y sekitar 28% saja, sedangkan regresi kudratik sekitar 46%
Bila kita menganggap hubungan X dan Y dalam polinom berderajat 3 atau polinomial kubik maka regresinya akan menghasilkan sebagaimana Gambar 3. Fungsi plonomial kubik mampu menjelaskan keragaman Y sekitar 52% saja.

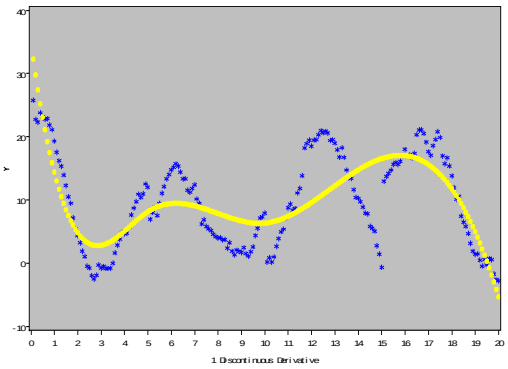


Gambar 3. Plot regresi parametrik polinomial kubik, R2= 0.52106.

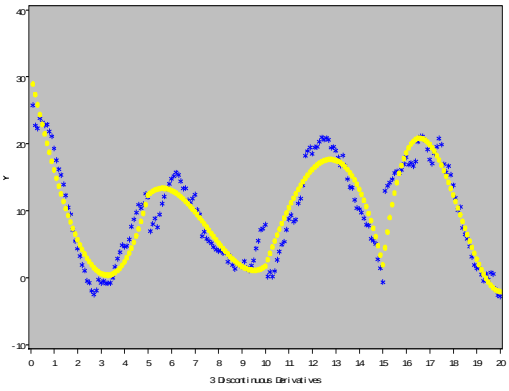
Gambar 2 dan 3 menunjukkan bahwa pendekatan parametrik gagal menduga fungsi f(x), karena tidak fleksibel. Berikut akan ditunjukkan bahwa pendekatan segmentasi pada pengamatan x mampu memberikan kepasan model yang lebih baik, melalui penempatan knot.

Regresi spline kubik dengan penempatan knot

Gambar 4 menunjukkan plot polinomial spline kubik yang merupakan fungsi jumlah terboboti dari satu fungsi konstan, satu fungsi linier garis lurus, kuadratik, dan kubik pada x<5. Fungsi polinomial kubik yang berbeda pada masing-masing bagian x, 5 < x < 10, 10 < x <15, dan 15<x. Fungsi spline ini lebih mulus dari fungsi kudratik namun lebih mendekati data sebenarnya daripada regresi polinom kuadratik, dengan R2=56.266%

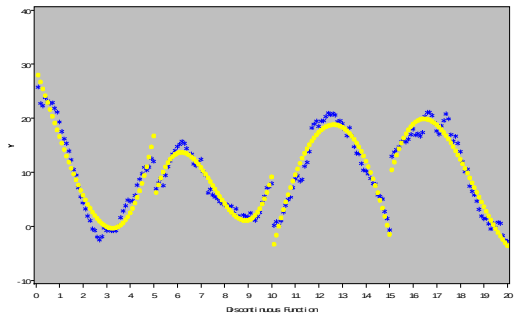


Gambar 4. Polinom tersegmen spline kubik dengan knot = 5, 10, 15 (R^2=56.266%).



Gambar 5. Polinom spline kubik tersegmen dengan knot = 5, 5, 5, 10, 10, 10, 15, 15, 15 (R2=0.95867).

Sedangkan bila kita dekati dengan polinom spline kubik tersegmen dengan knot = 5,5,5,10,10,10, 15,15,15 menghasilkan R^2 = 95.867%, garis spline sangat mendekati data, tetapi sedikit kurang mulus pada titik-titik knotnya (Gambar 5).



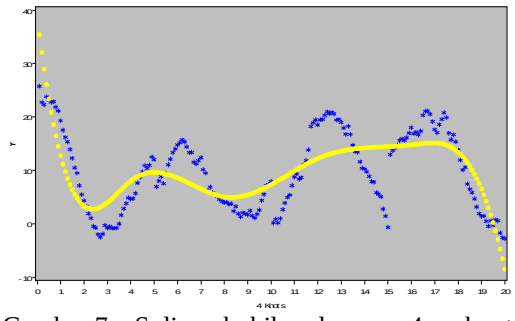
Gambar 6. Spline kubik takkontinu, knot = 55, 5, 5, 10, 10, 10, 10, 15, 15, 15, 15 (R2=0.98209).

Bila dengan spline kubik takkontinu dengan knot = 55, 5, 5, 10, 10, 10, 10, 15, 15 15, 15 (Gambar 6), sangat dekat dengan data (R2=98.209%), namun tidak halus dan takkontinu.

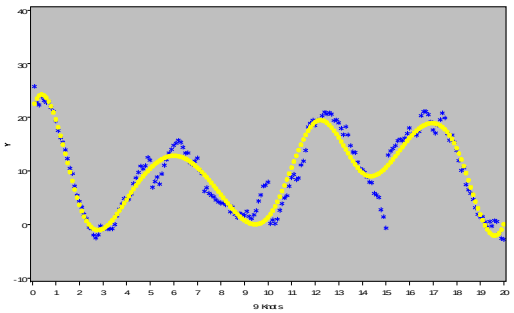
Regresi spline kubik dengan penempatan knot pada quantil

Dari hasil di atas dapat dikatakan bahwa kita tidak tahu secara pasti posisi knot yang memberikan segmentasi fungsi yang tepat pada titik ketakkontinuan. Dengan kata lain, dalam menduga fungsi $f(x)$ knot-knot tidak bisa ditempatkan disembarang tempat. Cara termudah dan umum digunakan adalah dengan menempatkan sejumlah knot yang telah ditetapkan, pada daerah-daerah yang sesuai dengan kuantil peubah penjelas *x*. fungsi spline yang dihasilkan adalah fungsi yang kontinu, dengan parameter pemulus/penghalus yang optimum menurut kriteria pemulus kuadrat tengah galat dan penalti kekasaran sehingga selain mulus juga akan cukup menggambarkan bentuk data.

Untuk itu perlu didefinisikan banyaknya knot yang digunakan.



Gambar 7. Spline kubik, dengan 4 knot (R2=0.69226).



Gambar 8. Polinomial spline kubik, dengan 9 knot (R2=0.94991).

Gambar 7 menunjukkan spline dengan empat buah knot pada kuantilnya. Spline dengan empat knot ternyata tidak bisa menggambarkan data dengan baik karena hanya mampu menerangkan sekitar 70% dari data.

Spline dengan sembilan knot ditempatkan pada desil (Gambar 8). Tampak bahwa spline yang dihasilkan hampir mendekati data sebenarnya dan mampu menerangkan keragaman peubah respon y sekitar 95% dan terlihat mulus disepanjang pengamatan *x*. Dari Gambar 7 dan 8 di atas jelas bahwa penempatan knot dan jumlah knot sangat mempengaruhi pemulusan regresi spline.

Regresi spline pada pengukuran tulang belakang manusia

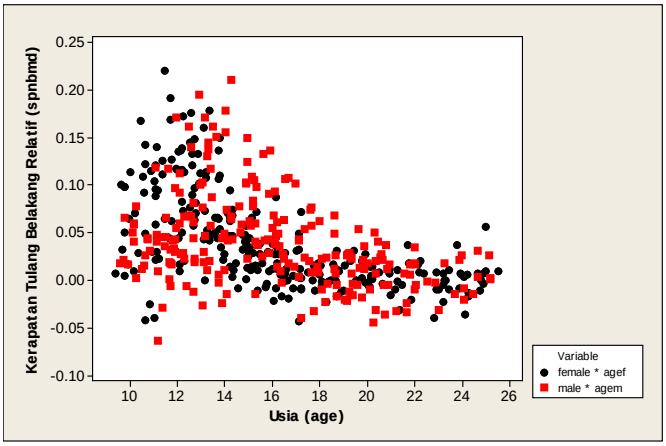
Hasil dari studi simulasi di atas diterapkan pada data pengukuran kerapatan relatif tulang belakang manusia dari 485 orang (226 perempuan dan 259 laki-laki) di Amerika Utara, tahun 1999 dengan peubah usia, menurut gender. Bila kita mencurigai terdapat pola yang berbeda antara laki-laki dan perempuan, kita dapat memutuskan untuk memodelkannya secara terpisah.

Pertama kita melihat pola tebaran data usia dengan kerapatan tulang menurut jenis kelamin, laki-laki dan perempuan (Gambar 9). Dari hasil regresi spline untuk masing-masing jenis kelamin (Tabel 1) didapat bahwa terdapat perbedaan nilai minimun pada *Generalized Cross-Validation* (GCV).

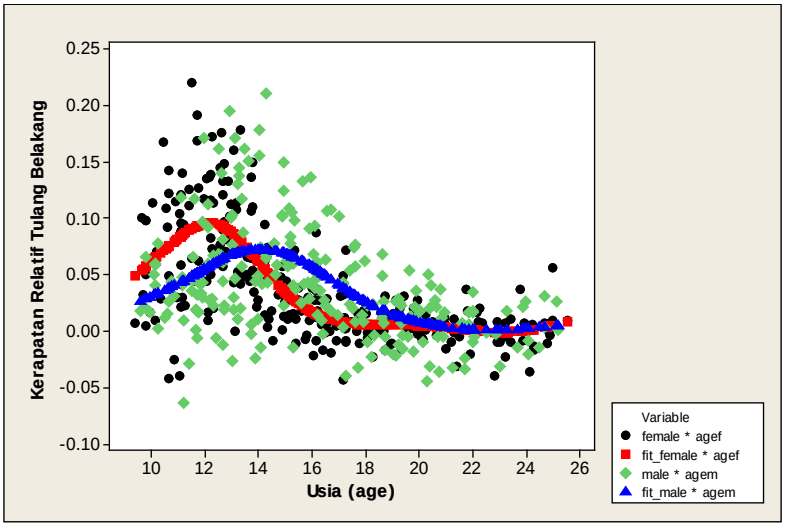
Tabel 1. Nilai GCV optimum untuk kedua model spline.

Perempuan		Laki-laki	
log10(n*Lambda)	GCV	log10(n*Lambda)	GCV
0	0.001291	1.4	0.001752
0.1	0.001287	1.41	0.001752
0.2	0.001283	1.42	0.001752
0.3	0.001279	1.43	0.001752
0.4	0.001276	1.44	0.001752
0.5	0.001273	1.45	0.001752
0.6	0.00127	1.46	0.001752*
0.7	0.001268	1.47	0.001752
0.8	0.001267	1.48	0.001752
0.9	0.001267*	1.49	0.001752
1	0.001267	1.5	0.001752

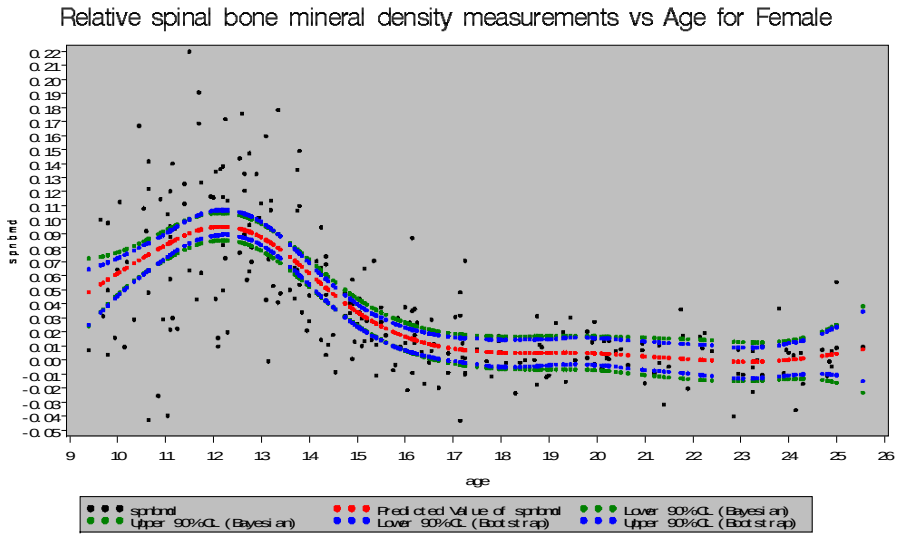
Keterangan: * mengindikasi nilai GCV minimum



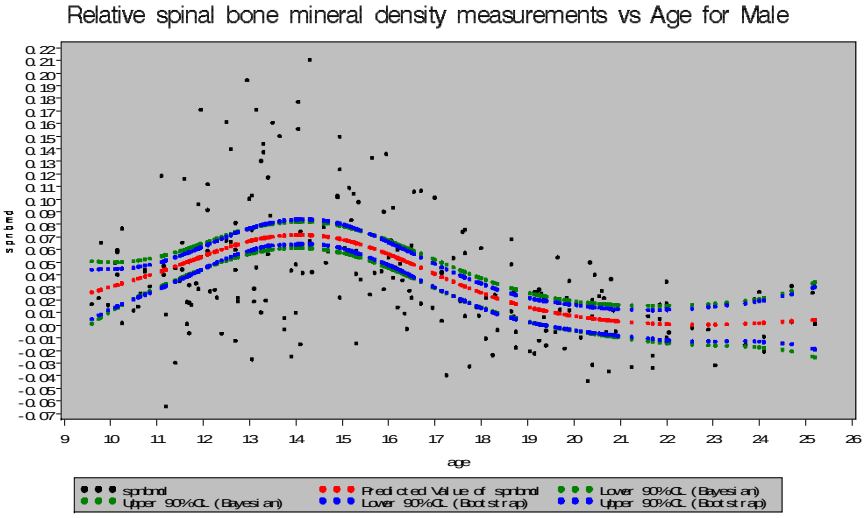
Gambar 9. Pola tebaran usia terhadap kerapatan tulang belakang menurut jenis kelamin, ●=perempuan, ■=laki-laki.



Gambar 10. Pengepasan regresi spline usia terhadap kepadatan tulang belakang, ●=perempuan, ▲=laki-laki.



Gambar 11. Selang kepercayaan bayesian dan bootstrap bagi regresi spline untuk perempuan (n=259).



Gambar 12. Selang kepercayaan bayesian dan bootstrap bagi regresi spline untuk laki-laki (n=226).

Tabel 2. Statistik pendugaan regresi spline.

Nilai Dugaan	Laki-laki	Perempuan
log10(n*Lambda)	0.9006	1.4644
Pinalti pemulus	0.0012	0.0003
Jumlah Kuadrat Sisa	0.3090	0.3767
Tr(I-A)	251.3630	220.4390
DF Model	7.6370	5.5610
Simpangan Baku	0.0351	0.0413

Hal ini diperkuat oleh hasil statistik pendugaan regresi spline (Tabel 2) bahwa terdapat perbedaan pada hasil pendugaan jumlah kuadrat sisa dan pinalti pemulus. Kepadatan relatif tulang belakang perempuan tampak mempunyai keragaman yang lebih rendah dibanding laki-laki. Hasil plot regresi disajikan pada Gambar 10.

Tampak bahwa pada perempuan kepadatan relatif lebih cepat menurun dengan bertambahnya usia, dibanding laki-laki. Pada usia-usia remaja telah terjadi penurunan sementara laki-laki baru mengalaminya pada awal usia 20 tahun.

Gambar 11 dan 12 memberikan selang penduga bagi dugaan titik regresi y=f(x) pada model kerapatan tulang belakang terhadap usia bagi laki-laki dan perempuan. Selang kepercayaan bayesian diperoleh berdasarkan Wahba (1983) dengan menggunakan prior galat yang menyebar normal dengan rata-rata nol. Sedangkan selang kepercayan bootstrap mengikuti Wang & Wahba (1995) dengan memberikan nilai $\hat{\sigma}^2$ dari data pada proses *bootstraping*.

Pada data kerapatan relatif ini tampak pula bahwa pada ukuran sample yang lebih kecil, dalam hal ini sample laki-laki, selang Bayes dan Bootstrap tidak begitu berbeda (Gambar 12). Sedangkan pada ukuran sample yang lebih besar (data perempuan), tampak selang bootstrap lebih sempit dari bayesian (Gambar 11).

KESIMPULAN

Regresi non parametrik spline merupakan regresi tersegmentasi yang memberikan keleluasaan pada fungsi polinomial yang berbeda pada tiap segmen dengan pemulus spline untuk memberikan kurva sepanjang pengamatan x. Untuk menduga fungsi f (x) sebaiknya menggunakan sejumlah knot yang ditetapkan dan ditempatkan pada daerah-daerah sesuai dengan kuantil peubah penjelas x sehingga didapat dugaan fungsi yang optimum pada perimbangan kuadrat tengah galat dan kemulusan. Dari sedangkan hasil penelitian mengenai selang kepercayaan Bayes dan Bootstrap tidak begitu berbeda. Pada ukuran sample yang lebih besar, tampak selang bootstrap lebih sempit dari bayesian.

Ucapan terimakasih

Penelitian ini didanai DIPA Unila Tahun Anggaran 2009 melalui Lembaga Penelitian Universitas Lampung

DAFTAR PUSTAKA

Eubank RL. 1988. *Spline smoothing and Nonparametrik Regression*. Marcel Dekker, Inc., New York.

Green PJ & Silverman BW. 1994. *Nonparametrik Regression and Generalized Linear Models (a roughness penalty approach*). Chapman & Hall, New York.

Hardle W. 1990. *Applied Non Paramertrik Regression.* Cambridge University Press, New York.

Hastie T, Tibshirani R & Freedman J. 2001. *The Elements of Statistical Learning: Data Mining, Inference,and Prediction.* Springer Verlag, New York.

Silverman BW. 1986. *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Scumacher L. 2007. *Spline Functions: Basic Theory.* 3Rd Ed. Vanderbilt University, Tennessee.

Takezawa K. 2006. *Introduction to Nonparametrics Regression.* John willey and Sons. USA.

Wahba G. 1990. *Spline Models for Observational Data*, SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wahba G. 1983. Bayesian "Confidence Interval" for the Craoss-validated Smoothing Spline. *J. R. Statist. Soc. B.* **45** (1): 133-150.

Wang Y & Wahba G. 1995. Bootstrap Confidence Intervals for Smoothing Splines and their Comparison to Bayesian Confidence Intervals. *J. Statistical Computation and Simulation*. **51**. [online abstract]

Wang J & Yang L. 2009. Polynomial Spline Sonfidence Bands for Regression Curves. *Statistica Sinica*. **19**: 325-342.